DOCUMENT RESUME

ED 159 226 TM 007 711

AUTHOR Reid, Jerry B.; Roberts, Dennis M.
TITLE A Monte Carlo Comparison of Phi and Kappa as Measures of Criterion-Referenced Reliability.
PUB DATE 27 Mar 78
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Analysis; *Correlation; *Criterion Referenced Tests; Cutting Scores; *Nonparametric Statistics; Simulation; Standard Error of Measurement; *Test Reliability
IDENTIFIERS *Kappa Coefficient; Monte Carlo Methods; *Phi Coefficient

ABSTRACT

Comparisons of corresponding values of phi and kappa coefficients were made for 270 instances of data generated by a Monte Carlo technique to simulate a test-retest situation. Data were generated for distributions with the same mean but three different levels of standard deviation, standard error of measurement and cutting score. Ten samples of 300 cases were used for each of the 27 crossed conditions. Comparisons of the corresponding phi and kappa values showed that none of the pairs of coefficients had differences larger than .04 and 98 percent had differences less than .01. Kappa is frequently recommended as the appropriate indicator for the reliability of criterion referenced tests, but since phi is a widely known correlation coefficient, its use may be preferable. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159226

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Monte Carlo Comparison of Phi and Kappa as Measures of Criterion-Referenced Reliability

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jerry B. Reid

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Jerry B. Reid

Dennis M. Roberts

The Pennsylvania State University

Presented at:

The American Educational Research Association

1978 Annual Meeting

Toronto, Canada

March 27, 1978

Printed in U.S.A.

TM007 711

Jerry B. Reid
Dennis M. Roberts
403 Carpenter Building
University Park, PA 16802

The reliability of a criterion-referenced test (CRT) has been defined as the consistency of decision making processes about mastery states. Kappa is frequently suggested in the literature as an indicator of the reliability of CRT's (Swaminathan, Hambleton, & Algina, 1974).

The formula for kappa appears very similar to that for phi (the special form of PPM for dichotomous data). Cohen (1960) noted this similarity and found that if the corresponding marginal proportions are the same, kappa will equal phi.

Although phi is more directly related to current correlational measurement techniques and familiar to more people, kappa seems to have been more widely accepted in the literature. The advocates of kappa have followed a line of reasoning which runs as follows: (1) there is potential restriction of variance in CRT's, (2) therefore classical measurement procedures which rely on variance are not appropriate, and (3) as a result, a new index--kappa--is proposed. The implicit assumption is that kappa will react differently to range restrictions than more conventional indices such as phi.

The purpose of this study was to compare the values of phi and kappa under several conditions of test score standard deviations, standard errors of measurement and cutting scores. It was hypothesized that none of the conditions would produce differences in the marginal proportions large enough to cause differences in the corresponding values of phi and kappa.

Method

A random sample generator from the Minitab system (Ryan, 1976) was used to simulate test scores on a test-retest paradigm under various conditions of population score standard deviation and standard error of measurement. The generating function operates in such a way that, once μ and σ are specified, a sample of given size will be randomly generated from a normal distribution having the stated parameters. The sample sizes for all cases reported here were N = 300.

Simulation Model. Normal distributions having $\mu = 15$ were assumed in each simulation. Population σ's took on values of 1, 2, and 3 representing various degrees of range restriction. Three values for the standard error of measurement (S_M) were used for each of the three population σ's. The arbitrary values used represented r_{XX} coefficients of .89, .56, and .00. Since S_M was calculated based upon the different σ's, the actual S_M values are different at each σ level (although the relative amount of measurement error is the same). In addition to the three levels each for σ and S_M, three different cutting scores for determining mastery and nonmastery were selected. These were 15, 17, and 19 corresponding to μ, 1 σ above μ (where $\sigma = 2$) and 2 σ units above μ (where $\sigma = 2$). The same

cutting scores were used for all cross comparison simulations. For 3 σ's, 3 S_M's, and 3 cutting scores, 3 x 3 x 3 = 27 cross comparisons resulted as shown in the appended Table 1.

Sampling. Ten N = 300 samples were randomly generated for each of the 27 crossed factor cells. Each sampling represented the following steps: (1) generation of a sample of N = 300 test scores with given μ and σ values, (2) estimating a true score for each observed X value, (3) setting up hypothesized observed score distributions using true scores and S_M as the μ and σ values, respectively, in the random sample generator, and (4) randomly selecting values from each hypothesized observed score distribution to be the retest scores.

Phi-Kappa Calculations. For each of the 270 sample simulations (27 x 10), the values for both phi and kappa were calculated based on their respective 2 x 2 contingency table. One dimension was first test score (master/nonmaster) and the other dimension was the retest score (master/nonmaster). Appended in Table 2 is a sample contingency table with both phi and kappa formulas.

Results

Of the 270 cases simulated, 30 phi-kappa pairs were indeterminate, i.e., the formulas yielded 0 over 0. Indeterminate values occurred when the combination of cutting score and population standard deviation was such that all examinees were classified into a single cell. Note that this also would be the case if the standard deviation equalled zero. Of the remaining 240 cases, none showed differences in the first decimal position. Approximately two percent had differences in the second decimal position. About two-thirds of the cases were different in the third and fourth decimal position. Appended in Figure 1 is a frequency polygon showing the discrepancies between phi and kappa.

All cases showing differences in the second decimal place came from contingency tables where the large majority of subjects were concentrated in the nonmaster-nonmaster cell. These contingency tables were, in turn, derived from conditions where the cutting score was at least 1 σ unit away from μ. The cases are not concentrated in any one r_{XX} or S_M condition.

Discussion

Considering the common practice of reporting reliability estimates to the second decimal position, only about seven percent of the differences between the phi-kappa pairs would show up in typical reporting procedures. Five of these seven percentage points are accounted for by rounding up from the third decimal place and would only produce differences of one unit in the second decimal position. Of the remaining two percent, only one case showed a difference as large as .04, the rest being only .02 or less. In addition to the finding that there are only a small percentage of "noticeable" differences in phi and kappa, in only one case did the difference occur at the highest r_{XX} value of .89 (phi-kappa = .02). Not only are the differences very small but almost all of those that do occur are at r_{XX} levels where differences of such magnitude are of little

practical importance. In summary, the reasons for making a choice between phi and kappa must be based upon personal preferences rather than concern for psychometric properties. However, since phi is more widely known and is, in fact, a correlational procedure, it is our opinion that the use of kappa should be dropped in favor of phi.

References

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Ryan, T. A. Minitab student handbook. Scitutate, Mass.: Duxbury Press, 1976.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Table 1

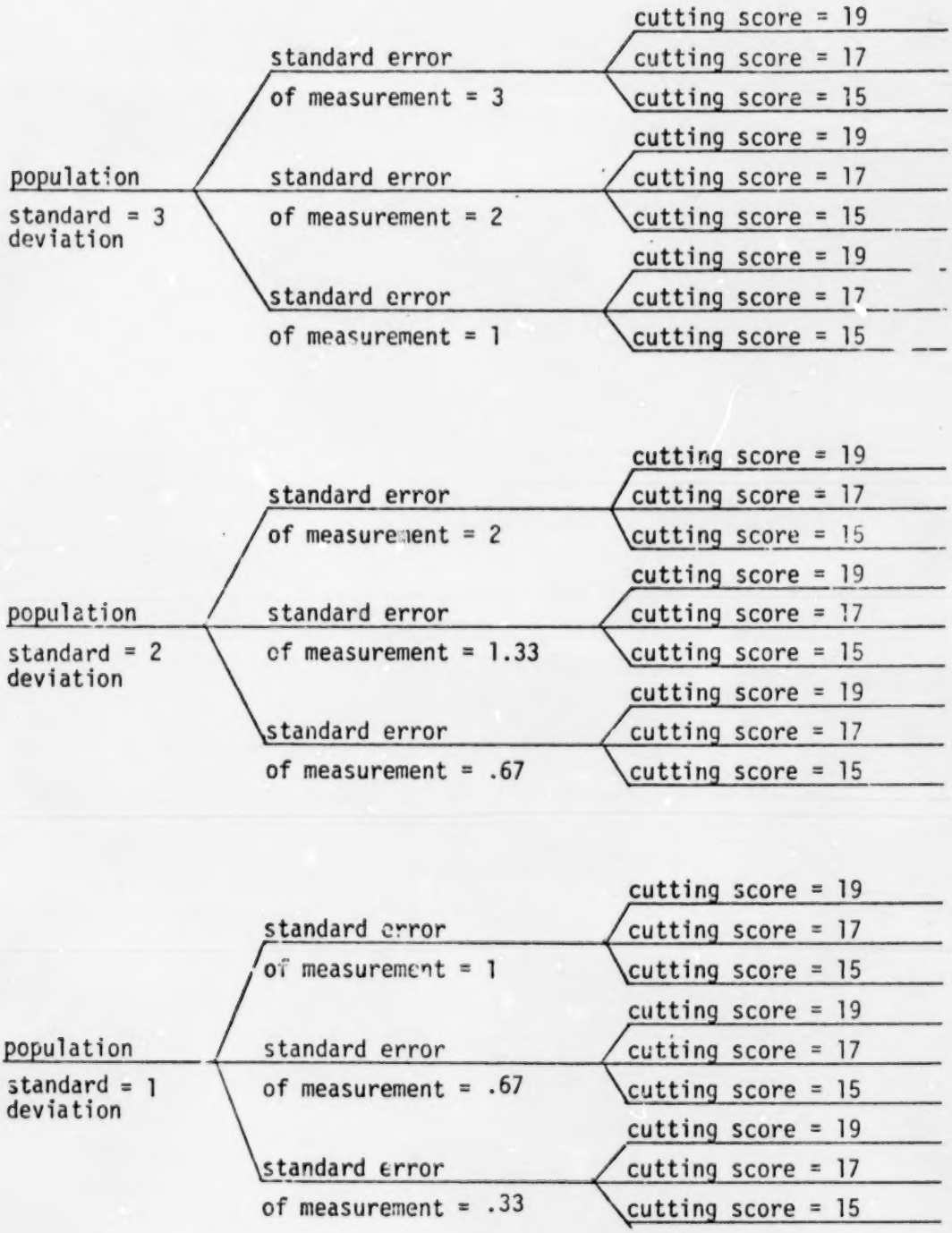
population
standard = 3
deviation
standard error
of measurement = 3
standard error
of measurement = 2
standard error
of measurement = 1
cutting score = 19
cutting score = 17
cutting score = 15
cutting score = 19
cutting score = 17
cutting score = 15
cutting score = 19
cutting score = 17
cutting score = 15
population
standard = 2
deviation
standard error
of measurement = 2
standard error
of measurement = 1.33
standard error
of measurement = .67
cutting score = 19
cutting score = 17
cutting score = 15
cutting score = 19
cutting score = 17
cutting score = 15
cutting score = 19
cutting score = 17
cutting score = 15
population
standard = 1
deviation
standard error
of measurement = 1
standard error
of measurement = .67
standard error
of measurement = .33
cutting score = 19
cutting score = 17
cutting score = 15
cutting score = 19
cutting score = 17
cutting score = 15
cutting score = 19
cutting score = 17
cutting score = 15

Table 2

Contingency Table

Test \ Retest	Master	Nonmaster
Master	A	B
Nonmaster	C	D

Note: A,B,C and D represent proportions.

Formula for Phi

$$\text{phi} = \frac{A \times D - B \times C}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Formula for Kappa

$$\text{kappa} = \frac{A + D - [(A+B)(A+C) + (C+D)(B+D)]}{1 - [(A+B)(A+C) + (C+D)(B+D)]}$$

Figure 1

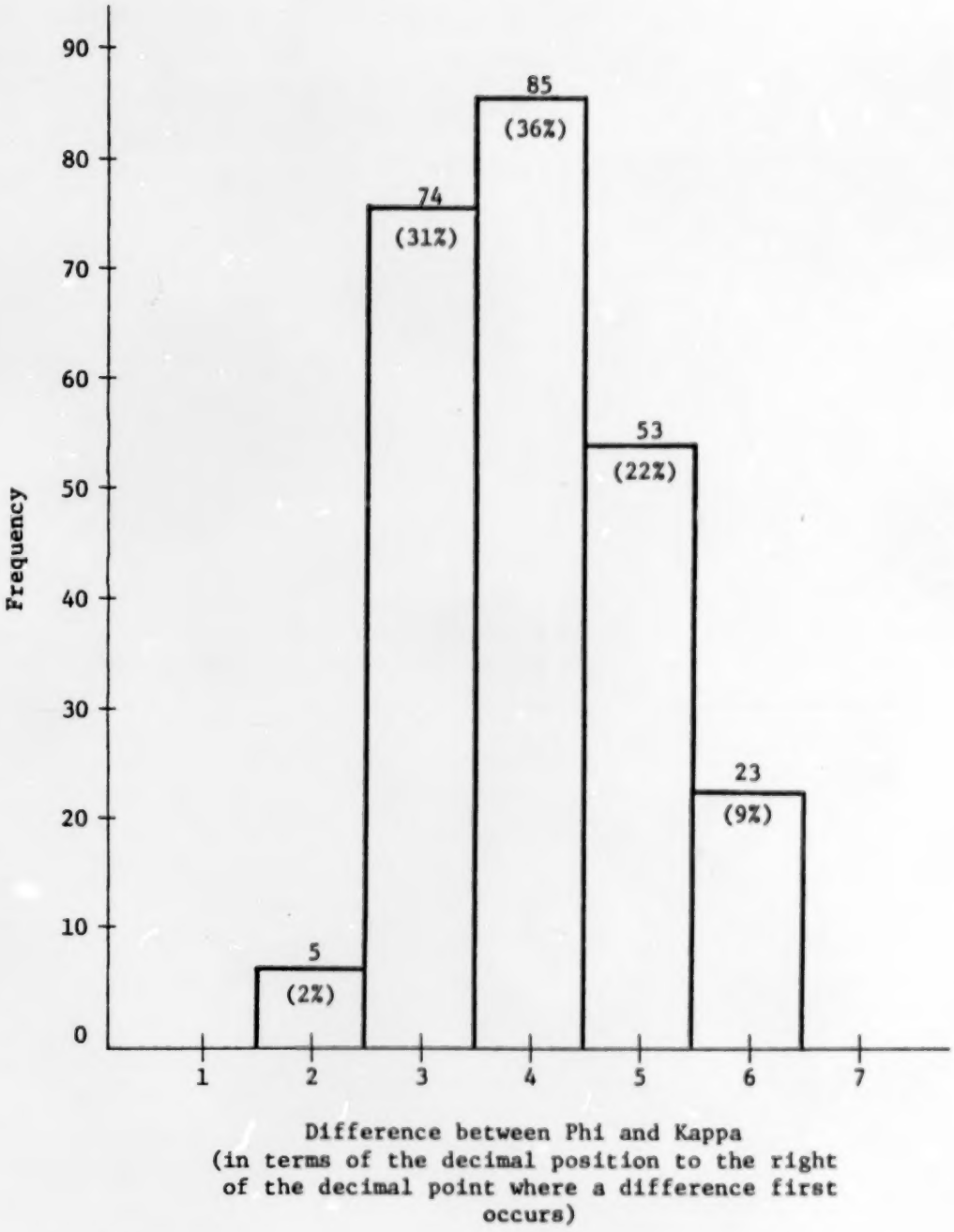

Figure 1. The Frequency of Differences of Various Magnitudes. (Note: Thirty cases were indeterminant.)